Biometrics 67, 1617–1626
December 2011
DOI: 10.1111/j.1541-0420.2011.01602.x

An Empirical Bayes' Approach to Joint Analysis of Multiple Microarray Gene Expression Studies

Lingyan Ruan* **and Ming Yuan****

H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332-0205, U.S.A.

**email:* lruan@gatech.edu

***email:* myuan@isye.gatech.edu

Summary. With the prevalence of gene expression studies and the relatively low reproducibility caused by insufficient sample sizes, it is natural to consider joint analysis that could combine data from different experiments effectively to achieve improved accuracy. We present in this article a model-based approach for better identification of differentially expressed genes by incorporating data from different studies. The model can accommodate in a seamless fashion a wide range of studies including those performed at different platforms by fitting each data with different set of parameters, and/or under different but overlapping biological conditions. Model-based inferences can be done in an empirical Bayes' fashion. Because of the information sharing among studies, the joint analysis dramatically improves inferences based on individual analysis. Simulation studies and real data examples are presented to demonstrate the effectiveness of the proposed approach under a variety of complications that often arise in practice.

Key words: Empirical Bayes'; Gene expression; Joint analysis; Mixture model.

1. Introduction

Microarray technology has presented unprecedented opportunities in genomic studies of complex diseases. It allows researchers to simultaneously monitor thousands of transcripts and discover novel biomarkers and genes. Despite their successes, these studies are often hampered by their relatively low reproducibility. This deficiency is often attributed to the high variability of gene expression measurements. Sources of distortion and noise are involved in almost every step along the process of taking gene expression measurements. It has long been recognized (e.g., Lee et al., 2000; Mukherjee et al., 2003) that such a problem could be alleviated through increased sample size. However, experiments with limited sample sizes remain common due to economic considerations. The recent explosion of popularity of high-throughput gene expression studies offers a more cost-effective alternative to this problem. With studies of the same diseases carried out independently by different research groups, it is natural to consider efficient ways of combining these data and jointly analyzing them. Through information sharing across studies, the accuracy of inferences could be greatly improved.

Because of its great potential, joint analysis of multiple experiments has attracted much attention in recent years. See, for example, Choi and Ghosh (2008) for a recent review. It is most commonly done through cross-experiment data normalization and transformation, which aims at translating and normalizing measurements from different sources on a common scale to allow for integration. In particular, Jiang et al. (2004) present a gene shaving method based on random forest (Breiman, 2001) and Fisher's linear discrimination analysis. Warnat, Eils, and Brors (2005) and Shabalin et al. (2008) also discuss different ways of integrating data through cross-experiment transformation. In general, however, it is difficult to integrate data without information loss and this would heavily bias each study. For example, van't Veer et al. (2002) and Wang et al. (2005) ended up with different predictive gene subsets with only three genes in common and there is no clear guidelines as to how it can be performed efficiently. Alternatively, one can also combine individual analysis results summarized by t-statistic, p-value, scored gene list, and so on (e.g., Rhodes et al., 2002; Choi et al., 2003; Ghosh et al., 2003; Parmigiani et al., 2004; Shen, Ghosh, and Chinnaiyan, 2004; Pyne, Futcher, and Skiena, 2006; Garrett-Mayer et al., 2007). In particular, Choi et al. (2003) propose to combine the effect size of genes from each study and conduct a permutation test to determine the significance level. Rhodes et al. (2002) and Pyne et al. (2006) consider ways of combining p-values of each study. Due to the small sample size of each study, the summary statistics obtained inevitably have high variations and subsequently these methods are subject to loss of efficiency in information sharing. This happens such as the studies of van't Veer et al. (2002) and Wang et al. (2005) mentioned above. It is also demonstrated by Mah et al. (2004) that detected genes on different platforms could have poor overlap. See Hong and Breitling (2008) for a comparison of methods and Rhodes et al. (2004), Parmigiani et al. (2004), and Scharpf et al. (2009) for other approaches.

There are also several major practical hurdles to joint analysis. In particular, there is no general consensus on how gene expression experiments should be conducted. As a result, the

choice of sample cohorts (e.g., age, ethnicity, and phase of disease), experiment platforms (e.g., cDNA or oligonucleotide), and processing facilities may all be different, and the scale of observations may not be comparable. These variations among experiments prohibit us from treating them as if they were simple replicates from a single study. In particular, a recent study in Kuo et al. (2002) compared Affymetrix and spotted cDNA and it was claimed that the correlation between the measurements from the two platforms was fairly low so it was unlikely that the two types of data could be transformed or normalized into a common standardized index. In practice, integrating multiple studies can be further complicated by missing data and sometimes, mismatch in biological conditions.

Consider, for illustration purpose, the study of prostate cancer, the most diagnosed cancer in men. There are a host of gene expression studies of prostate cancer. To motivate our work, microarray data were collected from four publicly available prostate cancer gene expression datasets generated independently by Dhanasekaran et al. (2001), Luo et al. (2001), Magee et al. (2001), and Welsh et al. (2001), respectively. One of the goals common to all four studies is in determining which genes are differentially expressed between locally advanced prostate cancer and benign tissue. The experiments, however, are done with different technologies. Dhanasekaran et al. (2001) and Luo et al. (2001) studies used spotted cDNA microarrays; whereas the other two experiments utilized Affymetrix GeneChips. In particular, the experiment from Magee et al. (2001) was conducted using HU6800 chip and Welsh et al. (2001) was done on U95A chip. Furthermore, these studies were performed on different but overlapping sets of genes. To overcome this problem, existing methods (see, e.g., Rhodes et al., 2002; Ghosh et al., 2003; Warnat et al., 2005) focus only on genes that are present in all studies. As we shall see in Section 4, such practice may result in more than 75% of the genes being discarded in some studies. Moreover, the remaining 25% of genes contain missing data, i.e., not all genes have complete observations from the samples tested. If the methods applied cannot allow missing data, this will reduce to only one gene (satisfying both intersection and complete data). This is clearly not an effective way of using the data. Another complication in combining the four experiments is the mismatch in biological conditions. Although all four studies include comparisons between locally advanced prostate cancer and benign prostate, Dhanasekaran et al. (2001) and Magee et al. (2001) also included a third biological condition: metastatic prostate cancer. Earlier attempts to combine these studies have either chosen to discard data collected from this condition or combine it with locally advanced cancer to form a new hypothesis.

These aforementioned limitations prompt us to develop a new technique. In this article, we propose a model-based method to integrate information from multiple experiments for the purpose of identifying differentially expressed genes among multiple biological conditions. Following Newton et al. (2001) and Kendziorski et al. (2003), we model the data from each individual study by a parametric empirical Bayes' model to share information across transcripts. These separate models are flexible to be applicable to different platforms and multiple biological conditions. Latent variables are then introduced to model the pattern of expression for a particular transcript and to share information across experiments. The modeling framework is fairly flexible and can handle a variety of practical issues including those mentioned above with ease.

The rest of this article is organized as follows. In the next section, we introduce the general modeling framework and show how statistical inferences can be efficiently conducted. Section 3 presents simulation studies to demonstrate the merits and versatility of the proposed method. We revisit the prostate cancer examples in Section 4 as well as another real data example before concluding with some remarks and discussions in Section 5.

2. Model and Inference

2.1 *Parametric Empirical Bayes' Model for a Single Study*

We begin with modeling gene expression data from a single study. Various methods have been developed for such purposes. Interested readers are referred to Parmigiani et al. (2003), Allison et al. (2006), and Do, Müller, and Vannucci (2006) for recent surveys. Here we adopt a parametric empirical Bayes' approach introduced by Newton et al. (2001) and Kendziorski et al. (2003).

Let x_{gcr} be the gene expression measurement taken from the rth replicate under condition c for gene g. Take the data from Dhanasekaran et al. (2001) as an example, three biological conditions ($c = 1$, 2, or 3), namely benign prostate, localized prostate cancer, or metastatic prostate cancer; 4839 genes ($g = 1, 2, \ldots, 4839$) are considered. A total of 14 replicates ($r = 1, 2, \ldots, 14$) are obtained for benign prostate; 14 for localized; and 20 for metastatic prostate cancer, respectively.

To fix ideas, we focus on two conditions ($c = 1$ or 2) in what follows. Sensible expression patterns concerning the comparison between two conditions for a particular gene include equivalent expression and differential expression. This can be formulated through latent variables μ_{gc} representing a population level of expression for gene g under biological condition c. Equivalent expression means that $\mu_{g1} = \mu_{g2}$ whereas differential expression indicates $\mu_{g1} \neq \mu_{g2}$. Our goal is therefore to infer such expression patterns from $\mathbf{x}_{g1\cdot} = (x_{g11}, x_{g12}, \ldots, x_{g1n_1})$ and $\mathbf{x}_{g2\cdot} = (x_{g21}, x_{g22}, \ldots, x_{g2n_2})$, where n_1 and n_2 are the number of replicates obtained under each condition, respectively. It is not hard to see that the marginal distribution of ($\mathbf{x}_{g1\cdot}$, $\mathbf{x}_{g2\cdot}$)

$$f(\mathbf{x}_{g1\cdot}, \mathbf{x}_{g2\cdot}) = (1-\pi) f(\mathbf{x}_{g1\cdot}, \mathbf{x}_{g2\cdot} \,|\, \mathrm{EE}) + \pi f(\mathbf{x}_{g1\cdot}, \mathbf{x}_{g2\cdot} \,|\, \mathrm{DE}), \tag{1}$$

where we use f to denote a generic density function, marginal, or conditional; and $\pi = P(\mathrm{DE})$. The two conditional distributions can be modeled through a two-level hierarchical model:

$$f\left(\mathbf{x}_{g1\cdot}, \mathbf{x}_{g2\cdot} \,|\, \mathrm{EE}\right) = \int \left\{ \prod_{k=1}^{n_1} f\left(x_{g1k} \,|\, \mu_{g1} = \mu; \theta\right) \right\} \times \left\{ \prod_{k=1}^{n_2} f\left(x_{g2k} \,|\, \mu_{g2} = \mu; \theta\right) \right\} f(\mu; \tau)\, d\mu; \tag{2}$$

$$f(\mathbf{x}_{g1\cdot}, \mathbf{x}_{g2\cdot} \,|\, \mathrm{DE}) = \int \left\{ \prod_{k=1}^{n_1} f(x_{g1k} \,|\, \mu_{g1}; \theta) \right\} \left\{ \prod_{k=1}^{n_2} f(x_{g2k} \,|\, \mu_{g2}; \theta) \right\} \times f(\mu_{g1}; \tau) f(\mu_{g2}; \tau)\, d\mu_{g1}\, d\mu_{g2}, \tag{3}$$

where θ and τ are parameters shared by all genes and determined by the experiment characteristics.

Two particular choices of $f(\cdot \,|\, \mu; \theta)$ and $f(\cdot; \tau)$ are advocated, often referred to as the lognormal–normal (LNN) model and gamma–gamma (GG) model. In the LNN model, $f(\cdot \,|\, \mu; \theta)$ is a lognormal distribution, i.e.,

$$f(x \,|\, \mu; \theta) = \frac{1}{\sqrt{2\pi\theta}} \exp\left\{ -\frac{(\ln x - \mu)^2}{2\theta} \right\}; \tag{4}$$

whereas $f(\cdot; \tau)$ is also a normal distribution with $\tau = (\tau_1, \tau_2)'$ represents the mean and variance parameter, respectively. Alternatively for the GG model, $f(\cdot \,|\, \mu; \theta)$ is a gamma distribution, i.e.,

$$f\left(x \,|\, \mu; \theta\right) = \frac{\lambda^\theta}{\Gamma(\theta)} x^{\theta-1} \exp\left(-\lambda x\right), \tag{5}$$

where the shape parameter is given by $\lambda = \theta/\mu$. $f(\cdot; \tau)$ is chosen such that λ also follows a gamma distribution

$$f(\lambda; \tau_1, \tau_2) = \frac{{\tau_2}^{\tau_1}}{\Gamma(\tau_1)} \lambda_g^{\tau_1 - 1} \exp\left(-\tau_2 \lambda\right). \tag{6}$$

Closed form expression are available for $f(\mathbf{x}_{g1\cdot}, \mathbf{x}_{g2\cdot} \,|\, \mu_{g1} = \mu_{g2})$ and $f(\mathbf{x}_{g1\cdot}, \mathbf{x}_{g2\cdot} \,|\, \mu_{g1} \neq \mu_{g2})$ with both LNN and GG models. The readers are referred to Kendziorski et al. (2003) for further details.

2.2 *Joint Modeling with Multiple Studies*

We now consider multiple studies. For brevity, we shall first assume that in each study, the same set of genes ($g = 1, 2, \ldots, G$) and the same set of conditions ($c = 1, 2, \ldots, C$) are considered. This assumption will later be relaxed. With slight abuse of notation, let $\mathbf{X}_s := \{x_{sgcr} : g = 1, \ldots, G;\ c = 1, \ldots, C;\ r = 1, \ldots, n_{sc}\}$ be the gene expression measurements obtained in the sth study ($s = 1, 2, \ldots, S$), where n_{sc} is the number of replicates under condition c in the study. Clearly $\mathbf{X}_s$ can be modeled using the parametric empirical Bayes' model discussed before. The hierarchical modeling can be summarized by the diagram below:

expression pattern (DE or EE)
↓
latent expression level $\{\mu_{sgc}\}$
↓
expression measurements $\{x_{sgcr}\}$

The latent expression levels are determined stochastically by the expression pattern through distribution $f(\mu; \tau)$ whereas the expression measurement by the latent levels through conditional distribution $f(x \,|\, \mu; \theta)$. Parameters θ and τ reflect the stochastic variation within a study and therefore are allowed to be experiment dependent. This is, in particular, necessary when handling studies from different platforms due to their difference in scales. To this end, we shall write θ_s and τ_s in what follows to emphasize the dependence between these parameters and the study. On the other hand, given that the same biological process is studied, a gene's differential expression pattern should remain the same across all studies.

Let $\mathbf{x}_{\cdot gc\cdot} = \{\mathbf{x}_{sgcr} : s = 1, \ldots, S;\ r = 1, \ldots, n_{sc}\}$ be the collection of all expression measurements obtained from all studies on gene g and condition c. Then the conditional distribution of these measurements under the two differential expression patterns can be given by

$$f\left(\mathbf{x}_{\cdot g1\cdot}, \mathbf{x}_{\cdot g2\cdot} \,|\, \mathrm{DE}\right) = \prod_{s=1}^{S} f\left(\mathbf{x}_{sg1\cdot}, \mathbf{x}_{sg2\cdot} \,|\, \mathrm{DE}\right); \tag{7}$$

$$f\left(\mathbf{x}_{\cdot g1\cdot}, \mathbf{x}_{\cdot g2\cdot} \,|\, \mathrm{EE}\right) = \prod_{s=1}^{S} f\left(\mathbf{x}_{sg1\cdot}, \mathbf{x}_{sg2\cdot} \,|\, \mathrm{EE}\right), \tag{8}$$

where the experiment specific conditional distributions are given in the previous subsection. In other words, for a randomly picked gene, its marginal distribution will follow a two-component mixture distribution:

$$f\left(\mathbf{x}_{\cdot g\cdot\cdot}\right) = \pi \prod_{s=1}^{S} f\left(\mathbf{x}_{sg1\cdot}, \mathbf{x}_{sg2\cdot} \,|\, \mathrm{DE}\right) + (1-\pi) \prod_{s=1}^{S} f\left(\mathbf{x}_{sg1\cdot}, \mathbf{x}_{sg2\cdot} \,|\, \mathrm{EE}\right), \tag{9}$$

where π is the probability that a randomly picked gene is differentially expressed. Note that from (9), data collected from different studies are not independent under our joint modeling framework because

$$f\left(\mathbf{x}_{\cdot g\cdot\cdot}\right) \neq \prod_{s=1}^{S} f\left(\mathbf{x}_{sg\cdot\cdot}\right),$$

where $f(\mathbf{x}_{sg\cdot\cdot})$ is the marginal density of the data collected from Study s as given by (1).

2.3 *Empirical Bayes' Inference*

If the experiment specific parameters θ_s and τ_s, $s = 1, \ldots, S$ are known, inference on a gene's expression pattern can be conducted through their posterior probabilities, i.e.,

$$P\left(\mathrm{DE} \,|\, \mathbf{x}_{\cdot g1\cdot}, \mathbf{x}_{\cdot g2\cdot}\right) = \frac{\pi f\left(\mathbf{x}_{\cdot g1\cdot}, \mathbf{x}_{\cdot g2\cdot} \,|\, \mathrm{DE}\right)}{\pi f\left(\mathbf{x}_{\cdot g1\cdot}, \mathbf{x}_{\cdot g2\cdot} \,|\, \mathrm{DE}\right) + (1-\pi) f\left(\mathbf{x}_{\cdot g1\cdot}, \mathbf{x}_{\cdot g2\cdot} \,|\, \mathrm{EE}\right)}, \tag{10}$$

where $\pi = P(\mathrm{DE})$ is the probability that a randomly selected gene is differentially expressed. According to Bayes' rule, we classify a gene as differentially expressed if the posterior probability of differential expression is greater than 50% and equivalent expression otherwise. These posterior probabilities

provide a natural means of inferring differential expression by integrating multiple studies.

Following Efron et al. (2001) and Newton et al. (2001), parameters $\{\theta_s, \tau_s\colon s = 1, \ldots, S\}$ as well as π can be estimated in an empirical Bayes' fashion. Note that these parameters are shared by all genes. The log-likelihood for all data can then be given by

$$\ell(\mathbf{x}_{\cdot 1\cdot}, \mathbf{x}_{\cdot 2\cdot}) = \sum_{g=1}^{G} \ell(\mathbf{x}_{\cdot g1\cdot}, \mathbf{x}_{\cdot g2\cdot}),$$

where

$$\ell(\mathbf{x}_{\cdot g1\cdot}, \mathbf{x}_{\cdot g2\cdot}) = \log\{(1-\pi) f(\mathbf{x}_{\cdot g1\cdot}, \mathbf{x}_{\cdot g2\cdot} \mid \mathrm{EE}) + \pi f(\mathbf{x}_{\cdot g1\cdot}, \mathbf{x}_{\cdot g2\cdot} \mid \mathrm{DE})\}.$$

The maximum likelihood estimator of all parameters θ_s and τ_s, $s = 1, \ldots, S$ and π can be efficiently computed using expectation-maximization (EM) algorithm by treating a gene's differential expression pattern (i.e., EE or DE) as missing.

Denote by z_g gene g's differential expression pattern. Then the log complete likelihood of parameter $\eta := \{\pi, \theta_s, \tau_s : 1 \le s \le S\}$ can be given as

$$\begin{aligned}\ell(\eta; \mathbf{x}_{\cdot\cdot\cdot\cdot}, z_{\cdot}) &= \sum_{g=1}^{G}\left[\mathbf{1}(z_g = \mathrm{DE})\left\{\log\pi + \sum_{s=1}^{S}\log f(\mathbf{x}_{sg\cdot\cdot} \mid \mathrm{DE}; \theta_s, \tau_s)\right\}\right.\\ &\qquad \left. + \mathbf{1}(z_g = \mathrm{EE})\left\{\log(1-\pi) + \sum_{s=1}^{S}\log f(\mathbf{x}_{sg\cdot\cdot} \mid \mathrm{EE}; \theta_s, \tau_s)\right\}\right].\end{aligned}$$

Let $\{\pi^{[t]}, \theta_s^{[t]}, \tau_s^{[t]} : 1 \le s \le S\}$ be the parameter estimates obtained from the tth iteration of the EM algorithm. Then in the $(t+1)$th iteration, we compute first the expectation of log complete likelihood with respect to z_gs given $\mathbf{x}$ and these parameter estimates:

$$\begin{aligned}Q(\eta) &= E_{Z_g:1\le g\le G}\,\ell(\theta; \mathbf{x}_{\cdot\cdot\cdot\cdot}, Z_{\cdot})\\ &= \sum_{g=1}^{G}\left[T_g\left\{\log\pi + \sum_{s=1}^{S}\log f(\mathbf{x}_{sg\cdot\cdot} \mid \mathrm{DE}; \theta_s, \tau_s)\right\}\right.\\ &\qquad \left. + (1-T_g)\left\{\log(1-\pi) + \sum_{s=1}^{S}\log f(\mathbf{x}_{sg\cdot\cdot} \mid \mathrm{EE}; \theta_s, \tau_s)\right\}\right],\end{aligned}$$

where

$$T_g = \frac{\pi^{[t]}\prod_{s=1}^{S} f\left(\mathbf{x}_{sg\cdot\cdot}, \mathbf{x}_{sg\cdot\cdot} \mid \mathrm{DE}; \theta_s^{[t]}, \tau_s^{[t]}\right)}{\pi^{[t]}\prod_{s=1}^{S} f\left(\mathbf{x}_{sg\cdot\cdot}, \mathbf{x}_{sg\cdot\cdot} \mid \mathrm{DE}; \theta_s^{[t]}, \tau_s^{[t]}\right) + (1-\pi^{[t]})\prod_{s=1}^{S} f\left(\mathbf{x}_{sg\cdot\cdot}, \mathbf{x}_{sg\cdot\cdot} \mid \mathrm{EE}; \theta_s^{[t]}, \tau_s^{[t]}\right)}.$$

In the second step, also called the M-step, we maximize Q with respect to θ to get an updated parameter estimation. In particular, it is clear that

$$\pi^{[t+1]} = \frac{1}{G}\sum_{g=1}^{G} T_g,$$

and (θ_s, τ_s) can be updated by the maxmizer of

$$Q_s(\theta_s, \tau_s) := \sum_{g=1}^{G}\{T_g \log f(\mathbf{x}_{sg\cdot\cdot} \mid \mathrm{DE}; \theta_s, \tau_s) + (1-T_g)\log f(\mathbf{x}_{sg\cdot\cdot} \mid \mathrm{EE}; \theta_s, \tau_s)\}.$$

In principle, a prior can also be assigned to these parameters and fully Bayesian inference can be made for the hierarchical model. We opt for the empirical Bayes' framework to avoid sophisticated and sometimes subjective prior elicitation.

2.4 *Missing Data*

As mentioned in Section 1, one of the most common difficulties associated with joint analysis is missing data. Due to limitations of technology and quality control, the set of genes measured in one dataset may not be the same as another dataset. In practice, only those genes measured across all experiments are included in the joint analysis. This can be a significant loss of information as we shall see in the prostate cancer data in Section 4, where 30% to 75% of the data from each experiment are wasted if this approach is taken. In contrast, this problem can be conveniently addressed within our framework. Rather than considering only genes that are present in all experiments, we include all genes that appears in at least one experiment. If a particular gene is not present in an experiment, we treat it as missing data. More specifically, let $\mathcal{M}_g (\subset \{1, \ldots, S\})$ be the collection of study indices where gene g is missing. Then the log complete likelihood becomes

$$\begin{aligned}\ell(\eta; \mathbf{x}_{\cdot\cdot\cdot\cdot}, z_{\cdot}) &= \sum_{g=1}^{G}\left[\mathbf{1}(z_g = \mathrm{DE})\left\{\log\pi + \sum_{s\notin\mathcal{M}_g}\log f(\mathbf{x}_{sg\cdot\cdot} \mid \mathrm{DE}; \theta_s, \tau_s)\right\}\right.\\ &\qquad \left. + \mathbf{1}(z_g = \mathrm{EE})\left\{\log(1-\pi) + \sum_{s\notin\mathcal{M}_g}\log f(\mathbf{x}_{sg\cdot\cdot} \mid \mathrm{EE}; \theta_s, \tau_s)\right\}\right].\end{aligned}$$

The EM algorithm proceeds in exactly the same fashion as before except that now the index s for the products and summations over studies now runs over $s \notin \mathcal{M}_g$ instead of $1 \le s \le S$ for gene G.

2.5 *Multiple Conditions and Condition Mismatch*

The proposed framework for joint analysis can be easily extended to handle more than two conditions. Consider, for example, the data taken from Dhanasekaran et al. (2001), where three biological conditions are investigated. For each condition, we introduce a latent gene expression level, μ_{sgc}, $c = 1$, 2 or 3. When comparing these conditions for gene g, we have the following equality or inequality conditions that may hold:

$$\begin{aligned}
&\text{Pattern 1:}\ \ \mu_{sg1} = \mu_{sg2} = \mu_{sg3},\\
&\text{Pattern 2:}\ \ \mu_{sg1} = \mu_{sg2} \neq \mu_{sg3},\\
&\text{Pattern 3:}\ \ \mu_{sg1} \neq \mu_{sg2} = \mu_{sg3},\\
&\text{Pattern 4:}\ \ \mu_{sg1} = \mu_{sg3} \neq \mu_{sg2},\\
&\text{Pattern 5:}\ \ \mu_{sg1} \neq \mu_{sg2} \neq \mu_{sg3}.
\end{aligned} \tag{11}$$

Similar to before, these latent expression level can be modeled by an experiment-specific distribution $f(\mu;\, \tau_s)$, where under pattern 1, all three latent expression levels are obtained as a single sample from $f(\cdot;\, \tau_s)$; under pattern 2, $\mu_{sg1} = \mu_{sg2}$ and μ_{sg3} are two independent samples from $f(\cdot;\, \tau_s)$ and so on. Similar formula as before can therefore be derived for $f(\mathbf{x}_{\cdot gc\cdot} \,|\, \text{Pattern } k)$:

$$\begin{aligned}
&f\left(\mathbf{x}_{\cdot g1\cdot}, \mathbf{x}_{\cdot g2\cdot}, \mathbf{x}_{\cdot g3\cdot} \,|\, \text{Pattern } k\right)\\
&= \prod_{s=1}^{S} f\left(\mathbf{x}_{sg1\cdot}, \mathbf{x}_{sg2\cdot}, \mathbf{x}_{\cdot g3\cdot} \,|\, \text{Pattern } k\right),
\end{aligned}$$

where the conditional densities can be computed and the inferences can also be conducted in a similar fashion as before.

A practical challenge that often arises with multiple biological conditions is the possible condition mismatch. Different experiments are designed to address and compare different but overlapping conditions. The overlap in biological conditions makes information sharing possible but the difference in biological conditions makes the information sharing difficult. For example, among the four prostate cancer studies we discussed earlier in the introduction, Dhanasekaran et al. (2001) considered three conditions including benign prostate, localized prostate cancer, and metastatic prostate cancer; whereas Luo et al. (2001) only investigated the first two conditions. A common practice is to ignore data obtained under the third condition from Dhanasekaran et al. (2001) and compare the first condition through a joint analysis. Although a convenient and sensible solution, it is clearly not the most efficient way of using data. In general, following this practice, when including multiple studies, we can only use those conditions that are present in all studies. Furthermore, as we shall demonstrate by simulations in the next section, doing so may result in loss of efficiency as well.

The problem of condition mismatch can also be handled conveniently within our proposed framework of joint analysis. For illustration purposes, we assume the one study has three conditions but the other one missed the third condition. With the third condition missing, it is evident that the expression measurements obtained from the second study have the following conditional distributions:

$$f(\mathbf{x}_{2g1\cdot}, \mathbf{x}_{2g1\cdot} \,|\, \text{Pattern } k) = \begin{cases} f(\mathbf{x}_{2g1\cdot}, \mathbf{x}_{2g1\cdot} \,|\, \text{EE}) & k = 1, 2\\ f(\mathbf{x}_{2g1\cdot}, \mathbf{x}_{2g1\cdot} \,|\, \text{DE}) & k = 3, 4, 5 \end{cases},$$

where $f(\mathbf{x}_{2g1\cdot}, \mathbf{x}_{2g1\cdot} \,|\, \text{EE})$ and $f(\mathbf{x}_{2g1\cdot}, \mathbf{x}_{2g1\cdot} \,|\, \text{DE})$ are defined by (2) and (3), respectively. The posterior probability of pattern k can therefore be evaluated as

$$\begin{aligned}
&P(\text{Pattern } k \,|\, \mathbf{x}_{\cdot g\cdot\cdot})\\
&= \frac{\pi_k f\left(\mathbf{x}_{1g\cdot\cdot} \,|\, \text{Pattern } k\right) f\left(\mathbf{x}_{2g\cdot\cdot} \,|\, \text{Pattern } k\right)}{\sum_{j=1}^{5} \pi_j f\left(\mathbf{x}_{1g\cdot\cdot} \,|\, \text{Pattern } j\right) f\left(\mathbf{x}_{2g\cdot\cdot} \,|\, \text{Pattern } j\right)}.
\end{aligned}$$

Parameter estimation can also be carried out in the same fashion as before.

3. Simulation Studies

3.1 *Benefit of Joint Analysis*

To demonstrate the effectiveness of the proposed method, we first conducted several sets of simulation studies. To demonstrate the benefit of joint analysis, we begin with a simple setting: two biological conditions, and no missing data. A total of $G = 5000$ genes and $S = 4$ experiments were simulated. For each experiment, $n_{sc} = 3$ replicates were simulated under each condition. The gene expression data were simulated from the LNN or GG model. Due to their similarity in performance, we report here only the results from LNN models. The simulation settings for each experiment are similar to those previously employed by Kendziorski et al. (2003) to mimic the real gene expression data and represent different experimental variations in practice. Denote $\eta = (\tau_1,\, \tau_2,\, \theta)$ the parameters associated with the LNN model. The parameters of the four experiments are set at $\eta_1 = (2,\, 0.5^2,\, 0.15^2)$, $\eta_2 = (5,\, 0.6^2,\, 0.25^2)$, $\eta_3 = (15,\, 1^2,\, 0.35^2)$, $\eta_4 = (30,\, 1.2^2,\, 0.45^2)$, respectively. These parameters are selected to mimic some of the main characteristics of the real data example to be presented later. Note that τ_2 reflects the variation of the latent mean of the gene expression levels such that larger values of τ_2 correspond to better separation between the two conditions for differentially expressed genes. In particular, the four studies used in this simulation set have average effect sizes of 1.62, 1.8, 2.64, and 3.23, respectively, where the average effect size is defined as the median of the effect sizes of differential expressions. A randomly chosen $\pi = 10\%$ genes are set to be differentially expressed. Both the joint and separate analyses were conducted. In the separate analysis, each experiment is analyzed separately using the empirical Bayes' approach of Kendziorski et al. (2003), referred to as EBarrays. We also apply the proposed approach for joint analysis. The operating characteristics of both analyses based upon 100 runs are summarized in Table 1.

We observe that joint analysis can significantly improve the separate analysis. Among the four experiments, experiment 4 has the strongest signal to noise ratio, which is also reflected by its superior performance to the other three experiments when analyzed separately. A possible misconception is that it is fruitless to combine such a good-quality experiment with others with relatively poor quality. Our result clearly suggests otherwise. It indicates that joint analysis can greatly improve even the experiment with the best quality.

To gain further insight of the merits of the proposed method, we now compare it with several alternative strategies of joint analysis. The first method is to naively combine the separate analysis of the four experiments by using the largest posterior probability of differential expression. The other two methods are taken from Choi et al. (2003) and Choi et al. (2007), respectively. Unlike the proposed method, these alternatives no longer connect with the posterior probability of differential expression and therefore it is unclear what a Bayes' rule means in these context. Nonetheless, each of these methods does provide a score, similar to the posterior probabilities, measuring the strength of evidence for differential expression. It is therefore of interest to know to what extent these scores

Table 1
Operating characteristics of joint analysis and separate analysis. The results are summarized from 100 *runs and all units are in percentages. The numbers in parentheses are the standard errors.*

		Separate analysis			
	Joint analysis	Experiment 1	Experiment 2	Experiment 3	Experiment 4
Sensitivity	99.34	53.16	58.39	78.87	91.9
	(0.037)	(0.245)	(0.258)	(0.185)	(0.109)
Specificity	99.99	99.48	99.49	99.78	99.92
	(0.001)	(0.013)	(0.012)	(0.007)	(0.004)
FDR	0.1	8.06	7.24	2.44	0.77
	(0.013)	(0.173)	(0.16)	(0.073)	(0.041)

can serve as proxies to identify differential expressed genes. A natural measure is the so-called area under the receiver operating characteristic curves (AUC). AUC, by definition, is necessarily between 0 and 1. The closer the AUC score is to one, the more effective the score is. For the proposed method and its three alternatives, the AUC, summarized from 100 runs are 99.66% (0.01%), 99.47% (0.02%), 60.98% (0.13%), and 83.41% (0.12%), respectively. The numbers in parentheses are the standard errors. Not surprisingly, the proposed method outperforms the alternatives because our simulation setting matches perfectly with the model settings.

To evaluate the robustness of the proposed method, we consider a more complex simulation set-up where the experimental data were generated as follows:

Experiment 1: The latent gene expression levels were simulated from an inverse gamma distribution with shape parameter 2 and location parameter 10. Then the gene expression measurements were simulated from a gamma distribution with the latent means and shape parameter 20.

Experiment 2: The latent means were simulated so that $A := \log((\mu_{2g1}\mu_{2g2})^{1/2})$ follows a uniform distribution between 5 and 11; and $M = \log(\mu_{2g1}/\mu_{2g2})$ follows a uniform distribution between -1 and 1 for differentially expressed genes and 0 for equivalently expressed genes. Then the observed gene expression measurements were simulated from gamma distribution with shape parameter 15.

Experiment 3: Similar to experiment 2 except that now M follows a uniform distribution between -2 and 2 and the expression measurements were simulated with shape parameter 25.

Experiment 4: Data were simulated from a LNN model with parameter $\theta = 0.3^2$ and $\tau = (2.3, 1.39^2)$.

The other settings are similar as before. The effect sizes of the four studies are 2.09, 1.65, 2.71, and 7.69, respectively. To gain further insights, we also consider three different percentages of differential expression: $\pi = 5\%$, 10%, and 20%. The sensitivity of joint four analysis ranges from 92% to 94% and individual analysis has sensitivity from 22% to 76%. It is evident that the joint analysis significantly outperforms separate analysis. We again compare the proposed method with the alternative methods for joint analysis in terms of AUC. Results are reported in Table 2. Similar to before, the proposed method enjoys superior performance.

Table 2
AUC comparison of the proposed method (*Proposed joint EB*), *and identifying differential expression based on largest posterior probability of separate analysis* (*Combined separate EB*) *and the methods of Choi et al.* (2003, 2007). *Results are based on* 100 *runs. All units are in percentages. The numbers in parentheses are standard errors.*

	$\pi = 5\%$	$\pi = 10\%$	$\pi = 20\%$
Proposed joint EB	99.63 (0.02)	99.61 (0.02)	99.59 (0.01)
Combined separate EB	98.56 (0.04)	98.48 (0.03)	98.49 (0.02)
Choi et al. (2003)	61.42 (0.19)	61.11 (0.14)	60.92 (0.12)
Choi et al. (2007)	57.94 (0.24)	58.86 (0.16)	60.15 (0.13)

3.2 *Missing Data*

We now consider the problem of missing data. To this end, we consider the following simulation scheme with a total of $G = 5000$ genes at two conditions. The proportion of DE genes is 5%. Similar to before, three replicates were simulated at every condition. Because of the robustness of the method, we focus here only on the LNN model with the parameters given before. The difference is now each experiment only involves a subset of the genes. In particular, experiment 1 includes 4500 randomly selected genes; and each of the remaining three experiments has 80% overlap with the first experiment and the set of overlapping genes is drawn randomly. In addition, experiments 2 and 3 each have 250 new genes randomly selected from the 500 genes not included in experiment 1. Experiment 4 covers all 500 genes not available in experiment 1. As a result, experiments 2 and 3 each have 3850 genes whereas experiment 4 comprises 4100 genes.

Table 3 summarizes the results from 100 simulation runs. It is clear that joint analysis dramatically improves the sensitivity with lower false discovery rate.

The methods of Choi et al. (2003, 2007) focus only on genes that are present in all studies. In the current setting, this results in discarding about half of the genes, which is clearly undesirable. Alternatively, the strategy of taking the largest posterior probability from separate analysis can still be applied, which yields an AUC of 93.07% (0.12%) based on 100 runs. This is to be combined with the proposed joint analysis which results in an AUC of 98.87% (0.04%).

Table 3

Performance comparison between joint and separate analysis when there are missing data. All units are in percentages. The numbers in parentheses are standard errors.

		Separate analysis			
	Joint analysis	Experiment 1	Experiment 2	Experiment 3	Experiment 4
Sensitivity	70.33	32.16	32.13	32.44	32.75
	(0.351)	(0.324)	(0.372)	(0.417)	(0.349)
Specificity	99.73	99.75	99.74	99.75	99.75
	(0.007)	(0.009)	(0.011)	(0.01)	(0.009)
FDR	6.78	12.86	12.96	12.39	12.62
	(0.157)	(0.365)	(0.462)	(0.422)	(0.419)

3.3 *Condition Mismatch*

Our final simulation study is designed to illustrate the effect of condition mismatch. We adopt a similar simulation set as before, with 5000 genes and four experiments. There are a total of three biological conditions but one condition is missing at each of the first three experiments. Specifically, the first experiment has three replicates under the first condition, three under the second condition, and none under the third condition. The second experiment has three replicates under each of the first and third conditions, but none under the second condition. The third experiment features three replicates under each of the second and third conditions, and none under the first condition. The last experiment has three replicates under each of the three conditions. As we pointed out earlier, such condition mismatch is a direct consequence of different biological hypothesis of interest. In experiment 1, our interest is in comparing the first two conditions. The goal is therefore to determine genes that are differentially expressed between these two conditions. Similarly, in experiment 2, we want to identify genes that are differentially expressed between the first and third conditions; and experiment 3, between the second and third conditions. In the last experiment, there are five possible patterns as we discussed before, all patterns except for pattern 1 can be identified as differential expression.

Given the different hypotheses, the natural question is whether or not a joint analysis of all four experiments can be beneficial. For example, for the "investigators" of the first experiment, combining with data on the first two conditions from the last experiment might be helpful, but it is not immediately clearly whether or not it helps if we include all four experiments. To illustrate the merits of the proposed joint analysis of all experiments, we apply three different strategies here: separate analysis of the first experiment; joint analysis of the first experiment and the last experiment with data from the third condition discarded; and the proposed method of joint analysis of all four experiments with missing conditions handled as missing data as we discussed before. Table 4 summarizes the operating characteristics of all three methods averaged over 100 runs. It is clear that both joint analyses improve upon the separate analysis with the proposed method outperforms the joint analysis with only two experiments. Similar comparisons were conducted from the angles of the "investigators" of experiments 2 and 3 and the results remain similar. Now consider the last experiment where the goal is identify differentially expressed genes among all three conditions. We compare the joint analysis that uses data from all four experiments and the individual analysis that only uses data from the last experiment. The results are also given in Table 4, which suggests that joint analysis gives superior performance. Note that existing methods for joint analysis are not designed to handle condition mismatch and are therefore not included in the comparison.

4. Real Examples

To further illustrate the merits of the proposed method, we now return to the prostate cancer examples discussed before. As mentioned earlier, four public microarray datasets generated independently by Dhanasekaran et al. (2001), Luo et al. (2001), Magee et al. (2001), and Welsh et al. (2001) were collected to determine genes that are differentially expressed between benign prostate and cancer tumors. As stated before, the data were generated with different platforms: Dhanasekaran et al. (2001) and Luo et al. (2001) employed spotted cDNA microarrays whereas the other two experiments utilized Affymetrix technology. All four studies include comparisons between locally advanced prostate cancer and benign prostate. Dhanasekaran et al. (2001) and Magee et al. (2001) also included a third biological condition: metastatic prostate cancer. A total of 13,474 unique genes are present in at least one of the experiments. There is, however, a severe mismatch in the set of genes measured among the four experiments with less than 10% (1322) of the genes presented in all four experiments. Table 5 summarizes some basic information of the data and gives the number of genes overlapped between the four experiments.

We ran the joint analysis both with the LNN and GG model and the results are similar. Therefore, we focus here on the results from the LNN model. Similar to the simulation study conducted before, there are two primary hypotheses concerning differential expression. The goal is to identify genes that are differentially expressed between either of the two types of cancer tumors and benign prostate. In other words, among the five expression patterns given in (11), we are interested in identifying genes in patterns 2, 3, 4, and 5 as opposed to pattern 1. Hereafter, we shall refer to genes with pattern 1 equivalently expressed genes; and the genes with other patterns differentially expressed genes. Similar to earlier studies (see, e.g., Choi et al., 2003), a large number of genes demonstrate significant difference between prostate cancer and benign prostate. To fix ideas, we focus on the top one hundred

Table 4
Performance comparison of separate analysis and joint analysis with condition mismatch. All units are in percentages.

	DE in conditions 1 and 2				DE in conditions 1 and 3		
	Sensitivity	Specificity	FDR		Sensitivity	Specificity	FDR
Exp 1	69.87 (0.188)	99.62 (0.011)	2.97 (0.08)	Exp 2	85.92 (0.136)	99.85 (0.006)	0.99 (0.037)
Exp 1 and 4	86.37 (0.135)	99.69 (0.009)	2.02 (0.056)	Exp 2 and 4	93.32 (0.093)	99.87 (0.005)	0.79 (0.033)
All exp	94.74 (0.086)	99.64 (0.008)	2.12 (0.048)	All exp	96.29 (0.067)	99.83 (0.006)	0.98 (0.037)
	DE in Conditions 2 and 3				DE among three conditions		
	Sensitivity	Specificity	FDR		Sensitivity	Specificity	FDR
Exp 3	88.01 (0.106)	99.88 (0.005)	0.79 (0.03)	Exp 4	64.22 (0.173)	99.31 (0.015)	4.11 (0.089)
Exp 3 and 4	94.3 (0.087)	99.89 (0.004)	0.67 (0.026)				
All exp	96.87 (0.063)	99.84 (0.006)	0.9 (0.034)	All exp	98.44 (0.04)	99.95 (0.003)	0.21 (0.014)

Table 5
Basic information of the four prostate cancer datasets. D—data from Dhanasekaran et al. (2001); *L—data from Luo et al.* (2001); *M—data from Magee et al.* (2001); *and W—data from Welsh et al.* (2001).

		Number of replicates			Pairwise overlap genes			
	Array type	Benign	Local PCA	Metastatic PCA	D	L	M	W
D	cDNA	14	14	20	4839	2642	1596	2126
L	cDNA	9	16	0		6109	2895	3574
M	Affy	4	8	3			5228	4963
W	Affy	9	23	0				9071

genes identified to follow patterns 2, 3, 4, or 5 by joint analysis. All of these genes have posterior probabilities of differential expression greater than 99%. Among these genes, 31 genes are not identified by any studies; 69 are identified to be differentially expression with posterior probability at least 95% in at least one of the fours studies when analyzing the four datasets separately; 34 in at least two studies; 7 in three studies; and 0 in all four studies. For the top 100 genes found by joint analysis, the Venn diagram in Figure 1 shows the availability in each of the individual analysis.

Joint analysis reveals significant genes agreed across studies more than would be expected by chance. Also, among the genes that are identified by joint analysis but appear not to be differentially expressed in individual analysis of each of the four studies is Hs.296638, a known prostate differentiation factor.

A second example we consider here is the four liver cancer datasets from Choi et al. (2003). All data were generated at two biological conditions: normal and tumor tissues. The goal is to identify genes differentially expressed in normal and tumor tissues. The datasets are of relatively poor quality when compared with the prostate datasets and have been used earlier in Choi et al. (2003) primarily to demonstrate the necessity of a joint analysis. Table 6 gives some basic information of the data and the number of genes that overlapped between the four experiments.

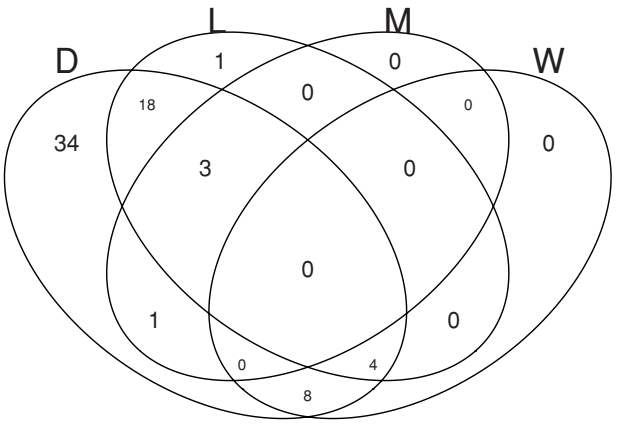

Figure 1. Venn diagram of differentially expressed genes for the four prostate cancer datasets: 100 genes are selected by joint analysis. Among them, 69 were identified by separate analysis on at least one dataset. This figure shows how many are identified by each one of the four datasets.

Similarly, the analysis is based on the LNN model. In joint analysis, the top 18 genes have posterior probability of differential equation greater than 90%. We evaluate the information of these genes and exclude genes of unknown functions. Then

Table 6
Basic information of the four liver cancer datasets

	Number of replicates		Pairwise overlap genes			
	Normal	Tumor	D	L	M	W
D1	16	16	10,314	10,289	10,194	9,921
D2	23	23		10,311	10,202	9,906
D3	29	5			10,216	9,815
D4	12	9				9,931

Table 7
In liver cancer, list of 10 *genes identified as differential expressed in joint analysis but identified by at most one study. The first gene failed to be selected by all studies.*

Tissue	Name
21.2.D.1	TATA box binding protein(TBP), mRNA
15.2.F.4	AL564975 cDNA
15.2.H.9	IL3-CT0219-271099-022-C02 cDNA
16.1.C.4	KIAA0107 gene product(KIAA0107), mRNA
19.4.D.12	KIAA0304 gene product(KIAA0304), mRNA
2.2.D.2	Thioredoxin reductase 1(TXNRD1), mRNA
20.2.A.9	Triosephosphate isomerase 1(TPI1), mRNA
22.3.A.4	Ribosomal protein L13a(RPL13A), mRNA
23.3.A.4	CD24 antigen (small cell lung carcinoma cluster 4 antigen) (CD24), mRNA
7.1.E.7	Hepatocyte growth factor regulated tyrosine kinase substrate (HGS), mRNA

we get 10 genes. In particular, 9 out of 10 are identified by only one study and 1 is not by any studies. Table 7 shows the information of these genes.

5. Conclusions

With the explosion of popularity of microarray experiments, it becomes a necessity to develop statistical methods that can effectively integrate data from multiple studies. Joint analysis of multiple experiments can alleviate the low sample size and high variability problem that is often faced in individual studies. In this article, we propose a model-based joint analysis of gene expression data from multiple studies to determine differentially expressed genes between multiple biological conditions. The proposed method shares information both among genes within one study and across studies without data transformation. The method is flexible to handle various practical complications such as missing data and condition mismatch. Simulation studies and real data examples show that the accuracy of statistical inferences can be drastically improved when using the proposed approach to combine multiple studies. It was demonstrated that combining data from multiple sources leads to increased sensitivity and specificity. Even incorporating those seemingly less optimal experiments could prove beneficial.

The aim of the proposed approach is to extract differential expression, to a certain degree, agreed upon by multiple studies. In addition to these identified genes, others that are not consensus choices sometimes may also be of interest. Discordance across studies is often observed in practice (see, e.g., Parmigiani et al., 2004). It may be attributed to faulty probe annotation (see, e.g., Dai et al., 2005) or hidden phenotypical heterogeneity or subtypes. Understanding such causes of discordance can also be of great importance.

Acknowledgements

The authors thank the editor, the associate editor, and two reviewers for their constructive comments that helped to greatly improve the presentation of the article. This research was supported in part by NSF grants DMS-0706724 and DMS-0846234, NIH grant R01GM076274-01, and a grant from the Georgia Cancer Coalition.

References

Allison, D. B., Cui, X., Page, G. P., and Sabripour, M. (2006). Microarray data analysis: From disarray to consolidation and consensus. *Nature Reviews Genetics* **7,** 55–65.

Breiman, L. (2001). Random forests. *Machine Learning* **45,** 5–32.

Choi, H. and Ghosh, D. (2008). A comparison of meta-analysis methods for gene expression data. In *Statistical Advances in the Biomedical Sciences*, A. Biswas, S. Datta, J. P. Fine, and M. R. Segal (eds), 200–215. New York: Wiley.

Choi, H., Shen, R., Chinnaiyan, A., and Ghosh, D. (2007). A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments. *BMC Bioinformatics* **8,** 364–383.

Choi, J. K., Yu, U., Kim, S., and Yoo, O. J. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* **19,** i84–i90.

Dai, M., Wang, P., Boyd, A., Kostov, G., Athey, B., Jones, E., Bunney, W., Myers, R., Speed, T., Akil, H., Watson, S., and Meng, F. (2005). Evolving gene/transcript definitions signicantly alter the interpretation of GeneChip data. *Nucleic Acids Research* **33,** e175.

Dhanasekaran, S. M., Barrette, T. R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pientas, K. J., Rubin, M. A., and Chinnaiyan, A. M. (2001). Delineation of prognostic biomarkers in prostate cancer. *Nature* **412,** 822–826.

Do, K. A., Müller, P., and Vannucci, M. (2006). *Bayesian Inference for Gene Expression and Proteomics*. Cambridge, U.K.: Cambridge University Press.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96,** 1151–1160.

Garrett-Mayer, E., Parmigiani, G., Zhong, X., Cope, L., and Gabrielson, E. (2007). Cross study validation and combined analysis of gene expression microarray data. *Biostatistics* **9,** 333–354.

Ghosh, D., Barette, T. R., Rhodes, D., and Chinnaiyan, A. M. (2003). Statistical issues and methods for meta-analysis of microarray data: A case study in prostate cancer. *Functional Integrative Genomics* **3,** 180–188.

Hong, F. and Breitling, R. (2008). A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics* **24,** 374–382.

Jiang, H., Deng, Y., Chen, H.-S., Tao, L., Sha, Q., Chen, J., Tsai, C.-J., and Zhang, S. (2004). Joint analysis of two microarray gene expression data sets to select lung adenocarinoma marker genes. *BMC Bioinformatics* **5**:81.

Kendziorski, C. M., Newton, M. A., Lan, H., and Gould, M. N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* **22,** 3899–3914.

Kuo, W. P., Jenssen, T.-K., Butte, A. J., Ohno-Machado, L., and Kohane, I. S. (2002). Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics* **18,** 405–412.

Lee, M.-L. T., Kuo, F. C., Whitmore, G. A., and Sklar, J. (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of the National Academy of Sciences of the United States of America* **97,** 9834–9839.

Luo, J., Duggan, D. J., Chen, Y., Sauvageot, J., Ewing, C. M., Bittner, M. L., Trent, J. M., and Isaacs, W. B. (2001). Human prostate cancer and benign prostatic hyperplasia: Molecular dissection by gene expression profiling. *Cancer Research* **61,** 4683–4688.

Magee, J. A., Araki, T., Patil, S., Ehrig, T., True, L., Humphrey, P. A., Catalona, W. J., Watson, M. A., and Milbrandt, J. (2001). Expression profiling reveals hepsin overexpression in prostate cancer. *Cancer Research* **61,** 5692–5696.

Mah, N., Thelin, A., Lu, T., Nikolaus, S., Kühbacher, T., Gurbuz, Y., Eickhoff, H., Klöppel, G., Lehrach, H., Mellgård, B., Costello, C.M., and Schreiber, S. (2004). A comparison of oligonucleotide and cDNA-based microarray systems. *Physiological Genomics* **16,** 361–370.

Mukherjee, S., Tamayo, P., Rogers, S., Rifkin, R., Engle, A., Campbell, C., Golub, T. R., and Mesirov, J. P. (2003). Estimating dataset size requirements for classifying DNA microarray data. *Journal of Computational Biology* **10,** 119–142.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, R. R., and Tsui, K. W. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8,** 37–52.

Parmigiani, G., Garett-Mayer, E., Irizarry, R., and Zeger, S. (2003). *The Analysis of Gene Expression Data: Methods and Software*. New York: Springer.

Parmigiani, G., Garrett-Mayer, E., Anbazhagan, R., and Gabrielson, E. (2004). A cross-study comparison of gene expression studies for the molecular classification of lung cancer. *Clinical Cancer Research* **10,** 2922–2927.

Pyne, S., Futcher, B., and Skiena, S. (2006). Meta-analysis based on control of false discovery rate: Combining yeast ChIP-chip datasets. *Bioinformatics* **22,** 2516–2522.

Rhodes, D. R., Barrette, T. R., Rubin, M. A., Ghosh, D., and Chinnaiyan, A. M. (2002). Meta-analysis of microarrays: Interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Research* **62,** 4427–4433.

Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T. R., Pandey, A., and Chinnaiyan, A. M. (2004). A large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 9309–9314.

Scharpf, R., Tjelmeland, H., Parmigiani, G., and Nobel, A. (2009). A Bayesian model for cross-study differential gene expression (with discussions). *Journal of the American Statistical Association* **104,** 1295–1323.

Shabalin, A. A., Tjelmeland, H., Fan, C., Perou, C. M., and Nobel, A. B. (2008). Merging two gene-expression studies via cross-platform normalizaiton. *Bioinformatics* **24,** 1154–1160.

Shen, R., Ghosh, D., and Chinnaiyan, A. M. (2004). Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. *BMC Genomics* **5**:94.

van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415,** 530–536.

Wang, Y., Klijn, J., Zhang, Y., Sieuwerts, A., Look, M., Yang, F., Talantov, D., Timmermans, M., Gelder, Meijer-van M., and Yu, J. (2005). Gene-expression profiles to predict distant metastasis of lymph-node negative primary breast cancer. *Lancet* **365,** 671–679.

Warnat, P., Eils, R., and Brors, B. (2005). Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics* **6**:265.

Welsh, J. B., Sapinoso, L. M., Su, A. I., Kern, S. G., Wang-Rodriguez, J., Moskaluk, C. A., Frierson, H. F., Jr, and Hampton, G. M. (2001). Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Research* **61,** 5974–5978.

Received June 2010. *Revised January* 2011.
Accepted February 2011.